INFORMATION AND CONTROL 4, 350–352 (1961)

Remark on the Paper by Bellman and Kalaba*

L. A. ZADEH

Dept. of Electrical Engineering, University of California, Berkeley 4, Calif.

The notion of an interrupted control process introduced in the paper by Bellman and Kalaba is a very significant one, since it substantially enlarges the class of control processes which can be treated by the techniques of dynamic programming.

By applying the principle of optimality, Bellman and Kalaba arrive at a functional equation with an implicit structure which, as they observe, is different from the usual functional equations of dynamic programming. The substance of our remark is that by using a somewhat different approach which is sketched in the sequel, one can reduce the problem of finding optimal policy for an interrupted stochastic control process to the same problem for a noninterrupted control process having a larger number of states. The simplification, however, is largely conceptual in nature, and we do not claim that it reduces computational labor.

More specifically, consider the same type of process as is treated in the paper by Bellman and Kalaba, and let $p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, y_t)$, $t = 0,1, \cdots, N-1$, denote the conditional distribution of $\mathbf{x}_{t+1}$,† the state at time $t+1$, given $\mathbf{x}_t$, the state at time t, and y_t, the input at time t. We assume that both $\mathbf{x}_t$ and y_t range over finite sets, $\mathbf{x}_t = \mathbf{q}_1, \cdots, \mathbf{q}_n$ and $y_t = \alpha_1, \cdots, \alpha_l$, $t = 0,1, \cdots, N$. The criterion function, C, is taken to be the expected value of a reward function h defined on the states at time N; i.e., $C = E\{h(\mathbf{x}_N)\}$. Furthermore, the probability of nonobservation of $\mathbf{x}_t$ at time t ($t = 0,1, \cdots, N$) is assumed to be a fixed constant p.

Let us enlarge the state space of the process by adding to the set of states $\{\mathbf{q}_1, \cdots, \mathbf{q}_n\}$, all states of the form $(\mathbf{q}_i ; \alpha_{i_1}, \alpha_{i_2}, \cdots, \alpha_{i_k})$ $1 \leqq k \leqq N-1$, $i = 1, \cdots, n$, where the symbol $(\mathbf{q}_i ; \alpha_{i_1}, \cdots, \alpha_{i_k})$

* This issue, pp. 346–349.

† For simplicity of notation, the same symbol is used to denote a random variable and its value.

means that $\mathbf{q}_i$ is the last state observed since the inputs $\alpha_{i_1}, \cdots, \alpha_{i_k}$ were applied to the system.

Let $\boldsymbol{\pi}_t$ denote a variable ranging over the enlarged state space. We shall refer to $\boldsymbol{\pi}_t$ as the *process state*, to distinguish it from the system state $\mathbf{x}_t$. In contrast to $\mathbf{x}_t$, which may or may not be observable, $\boldsymbol{\pi}_t$ is an observable variable.

It is evident that an interrupted stochastic control process in terms of the $\mathbf{x}_t$ is a noninterrupted stochastic control process in terms of the $\boldsymbol{\pi}_t$. Furthermore, for any given input sequence $y_0, y_1 \cdots y_{N-1}$, the $\boldsymbol{\pi}_t$ form a finite Markov chain in the sense that

$$p(\boldsymbol{\pi}_{t+1} \mid \boldsymbol{\pi}_t, y_t) \equiv p(\boldsymbol{\pi}_{t+1} \mid \boldsymbol{\pi}_t, \cdots, \boldsymbol{\pi}_0, y_t, \cdots, y_0) \tag{1}$$

The Markovian character of the $\boldsymbol{\pi}_t$ implies that the class of policies defined by the relation

$$y_t = D_t(\boldsymbol{\pi}_t), \tag{2}$$

where D_t is a function from the space of process states to the space of inputs, is complete in the sense that there is no better policy outside of this class.

It remains to be shown how the probability distribution $p(\boldsymbol{\pi}_{t+1} \mid \boldsymbol{\pi}_t, y_t)$ can be obtained from the knowledge of $p(\mathbf{x}_{t+1} \mid \mathbf{x}_t, y_t)$. Consider first the simple case where $\boldsymbol{\pi}_t = q_i$ and $\boldsymbol{\pi}_{t+1} = q_j$, with the input being $y_t = \alpha_k$. Clearly,

$$\Pr\{\boldsymbol{\pi}_{t+1} = \mathbf{q}_j \mid \boldsymbol{\pi}_t = \mathbf{q}_i, y_t = \alpha_k\} = (1-p)\Pr\{\mathbf{x}_{t+1} = \mathbf{q}_j \mid \mathbf{x}_t = \mathbf{q}_i, y_t = \alpha_k\} \tag{3}$$

with the right-hand member being known. Similarly,

$$\Pr\{\boldsymbol{\pi}_{t+1} = (\mathbf{q}_j; \alpha_{j_1}, \cdots, \alpha_{j_{k+1}}) \mid \boldsymbol{\pi}_t = (\mathbf{q}_j; \alpha_{j_1}, \cdots, \alpha_{j_k}), y_t = \alpha_{j_{k+1}}\} = p. \tag{4}$$

For more general transitions, we note that the conditional probability that the system is in state $\mathbf{q}_j$ at time t given that it is in process state $\boldsymbol{\pi}_t = (q_i; \alpha_{i_1}, \cdots, \alpha_{i_k})$ at time t is given by

$$\begin{aligned}
&\Pr\{\mathbf{x}_t = \mathbf{q}_j \mid \boldsymbol{\pi}_t = (\mathbf{q}_i; \alpha_{i_1}, \cdots, \alpha_{i_k})\} = \\
&\Pr\{\mathbf{x}_t = \mathbf{q}_j \mid \mathbf{x}_{t-k} = \mathbf{q}_i, y_{t-k} = \alpha_{i_1}, \cdots, y_{t-1} = \alpha_{i_k}\} \\
&\quad = \sum_{\mu} \cdots \sum_{\nu} \Pr\{\mathbf{x}_{t-k+1} = \mathbf{q}_\mu \mid \mathbf{x}_{t-k} = \mathbf{q}_i, y_{t-k} = \alpha_{i_1}\} \cdots \\
&\qquad \Pr\{\mathbf{x}_t = \mathbf{q}_j \mid \mathbf{x}_{t-1} = \mathbf{q}_\nu, y_{t-1} = \alpha_{i_k}\},
\end{aligned} \tag{5}$$

where all the terms on the right are known.

Then

$$\Pr\{\pi_{t+1} = \mathbf{q}_j \mid \pi_t = (\mathbf{q}_i\,;\,\alpha_{i_1}, \cdots, \alpha_{i_k}),\ y_t = \alpha_{i_{k+1}}\}$$
$$= (1-p) \sum_{\mu} \Pr\{\mathbf{x}_t = \mathbf{q}_\mu \mid \pi_t = (\mathbf{q}_i\,;\,\alpha_{i_1}, \cdots, \alpha_{i_k})\} \tag{6}$$
$$\times \Pr\{\mathbf{x}_{t+1} = \mathbf{q}_j \mid \mathbf{x}_t = \mathbf{q}_\mu,\ y_t = \alpha_{i_{k+1}}\}$$

with $\Pr\{\mathbf{x}_t = \mathbf{q}_\mu \mid \pi_t = (\mathbf{q}_i\,;\,\alpha_{i_1}, \cdots, \alpha_{i_k})\}$ being given by (5).

Thus, in this very straightforward fashion one can determine the transition probabilities $p(\pi_{t+1} \mid \pi_t,\ y_t)$ for the $\{\pi_t\}$ process from the transition probabilities $p(\mathbf{x}_{t+1} \mid \mathbf{x}_t,\ y_t)$ for the $\{\mathbf{x}_t\}$ process. Note that the Markovian property of the $\{\pi_t\}$ process expressed by (1) follows from (3)–(6) and the Markovian property of the $\{\mathbf{x}_t\}$ process.

To complete the reduction of the interrupted stochastic control process to a noninterrupted stochastic control process it is necessary to extend the definition of the function $h(\mathbf{x}_N)$ to a function $h^*(\pi_N)$ defined on the process states π_N. Here the underlying assumption is that the reward $h(\mathbf{x}_N)$ is determined by the actual state of the system at time N, but it is π_N and not necessarily $\mathbf{x}_N$ that is observable at time N. Under this assumption the reward associated with π_N will be the expectation of $h(\mathbf{x}_N)$ with respect to the conditional distribution $p(\mathbf{x}_N \mid \pi_N)$:

$$h^*(\pi_N) = \sum_{\mathbf{x}_N} h(\mathbf{x}_N)\ p(\mathbf{x}_N \mid \pi_N). \tag{7}$$

In this way, the problem of interrupted stochastic control of a system with state space $\{\mathbf{q}_1, \cdots, \mathbf{q}_n\}$ and reward function $h(\mathbf{x}_N)$ is reduced to the conventional problem of noninterrupted stochastic control of a system with state space $\{\mathbf{q}_1, \cdots, \mathbf{q}_n, (\mathbf{q}_i\,;\alpha_1), \cdots, (\mathbf{q}_n, \alpha_{i_1}, \cdots, \alpha_{i_{N-1}})\}$ and reward function $h^*(\pi_N)$ given by (7). The same technique can be employed in cases in which what is observed is not the state $\mathbf{x}_t$ but an output v_t, with the defining conditional distribution being of the form $p(\mathbf{x}_{t+1},\ v_t \mid \mathbf{x}_t,\ y_t)$. Furthermore, the possibility of nonobservation of v_t may depend on $\mathbf{x}_t$ and y_t instead of being a constant. Only minor modifications in the procedure sketched above are needed to reduce the problem of interrupted stochastic control for systems of this more general type to a corresponding problem for systems in which the process state is observable at each stage of the process.

RECEIVED: June 16, 1961